Home Work 3 (Case Study 1) – Collecting, Manipulating and Blending Data from Twitter

DS501 - Introduction to Data Science

Introduction

- Go to https://dev.twitter.com/apps/new and log in, if necessary
- Enter your Application Name, Description and your website address.
- Set the callback URL http://127.0.0.1:1410
- Accept the TOS, and solve the CAPTCHA.
- Submit the form by clicking the Create your Twitter Application
- Copy the consumer key (API key) and consumer secret from the screen into your application
- Download twitter package from https://github.com/geoffjentry/twitteR

Problem 1: Sampling Twitter Data with Streaming API about a certain topic

- Select a topic that you are interested in, for example, "#WPI" or "#DataScience"
- Use Twitter Streaming API to sample a collection of tweets about this topic in real time. (It would be recommended that the number of tweets should be larger than 50, but smaller than 500.
- Store the tweets you downloaded into a local file (csv file)

```
library(twitteR)
library(stringr)
setup_twitter_oauth(consumerKey, consumerSecret, accessToken, accessTokenSecret)
tweets = searchTwitter('#rstats', n=50)
tweetsDF = twListToDF(tweets)
```

Report some statistics about the tweets you collected

- The topic of interest: < INSERT YOUR TOPIC HERE>
- The total number of tweets collected: < INSERT THE NUMBER HERE>

Problem 2: Analyzing Tweets and Tweet Entities with Frequency Analysis

1. Word Count:

- Use the tweets you collected in Problem 1, and compute the frequencies of the words being used in these tweets.

```
# Your R code here
```

- Display a table of the top 30 words (ONLY) with their counts

```
#  Your R code here
```

2. Find the most popular tweets in your collection of tweets

- Please display a table of the top 10 tweets (ONLY) that are the most popular among your collection, i.e., the tweets with the largest number of retweet counts.

```
# Your R code here
```

3. Find the most popular Tweet Entities in your collection of tweets

Please display a table of the top 10 hashtags (ONLY), top 10 user mentions (ONLY) that are the most popular in your collection of tweets.

```
# Your R code here
```

Problem 3: Getting any 20 friends and any 20 followers of a popular user in twitter

- Choose a twitter user who has many followers, such as @hadleywickham.
- Get the list of friends and followers of the twitter user.
- Display 20 out of the followers, Display their ID numbers and screen names in a table.
- Display 20 out of the friends (if the user has more than 20 friends), Display their ID numbers and screen names in a table.
- Compute the mutual friends within the two groups, i.e., the users who are in both friend list and follower list, Display their ID numbers and screen names in a table

Problem 4 (Optional): Explore the data

Run some additional experiments with your data to gain familiarity with the twitter data and twitter API

Done

All set!

What do you need to submit?

Report: please prepare a document, preferably using RMarkdown (less than 10 pages) to report what you found in the data.

- What data you collected?
- Why this topic is interesting or important to you? (Motivations)
- How did you analyze the data?
- What did you find in the data? (please include figures or tables in the report)

Please create an R Markdown PDF including the R code in a report format.

How to submit: - Submit on Course Webpage on Canvas as a PDF file ONLY. Do not email it to me.